Bioinformatics

WAPAM

User manual

Index

1. Wapam, a weighted automata pattern matching....3
2Tutoriel : simple example of use....3
With Rdisk....3
Without Rdisk....6
3Input and output Data....7
Input data....7
Form « Multiple Lauch »....7
Ouput data....8
Web Format (HTML)....8
XML Format....9
CSV Format....9
Notice about the number of results....10
4Some details about Wapam functionalities....10
Weighted automata (WFA)....10
Wapam and Wapam/Rdisk....12
Performances....13
Specific needs....13
References....14

Illustrations Index

Illustration 1: example of input data filled in the Web interface.....4
Illustration 2: an other example of input data filled in the Web interface : search within a data base. 4
Illustration 3: an other example of input data filled in the Web interface : search within a genome...5
Illustration 4: an other example of input data filled in the Web interface : search within a personnal data bank.....5
Illustration 5: progression of the compilation of the Rdisk (FPGA) processors before the sequences filtration.....5
Illustration 6 : the results for the HTML example.....6
Illustration 7: positioning of the job launched in the genocluster tasks queue.....6
Illustration 8: example of the output data in HTML format with the option « each sequence matched».....8
Illustration 9: example of results with the option “each match”. In this sequence, the pattern appears twice at the positions 481 and 593.....9
Illustration 10: example of the output data in XML format 9
Illustration 11 : example of the output data in CVS format 10
Illustration 12 : a wheighted autamata of D-[ILV]-x(1,3)-A. 11
Illustration 13 : example of automata representing a Prosite pattern D-[ILV]-x(1,3)-A.....11
Illustration 14 : example of automata modified 12
Illustration 15 : WAPAM Material Achitecture.....12
Illustration 16 : comparison between the search times of pattern (*: estimates).....13

1 . Wapam, a weighted automata pattern matching

Wapam has been developped by the SYMBIOSE research team and OUEST-genopole® plateform offer an online version on their Web site. Wapam is a fast nucleic and proteic pattern matching tool, with or without errors, against complete genomes, data banks and personnal data banks.

The Web interface allows the users to execute the program on a backbone cluster (genocluster) supply by the plateform or to execute the program on a material accelerating Rdisk. Rdisk est a specialized structure realized by the research team SYMBIOSE to reduce the research time of the patttern in the sequences.

The first particularity of Wapam is that it treats pattern translated to weighted automata (WFA) (chapter 4). Weighted automata could be generated by Prosite pattern.

Each sequence goes along the automata. It return a score / threshold that allow to evaluate the adequacy between the sequence and the pattern. In fact, a score is the substitution errors number. If the score is higher than the threshold fixed, the pattern is detected at the current postition (for example, if one substitution is tolerated, the threshold is equal to -1 and the pattern is detected if the score is equal to or highter than -1. A Wapam research, with or without errors take the same execution time.

The other particularity of wapam is the connecting with the prototype Rdisk machine that allow a material accelerating of the computation. During a compilation stage, the pattern automata is translated to a specialized circuit. Each of the 31 processors that Rdisk is composed are parameterized with the circuit. The sequence is divided by 31 pieces that are analysed in each processor.

2 Tutoriel : simple example of use

With Rdisk

We would search the prosite pattern D-[ILV]-x(1,3)-A in a proteic data base SuissProt. We must generating an automata, clicking on « generate automata ».The automata is visible illustration 1 It's very important to be aware of that if a modification is done and the automata is already generated, you have to regenerate it. In this example, we decide to use Rdisk.

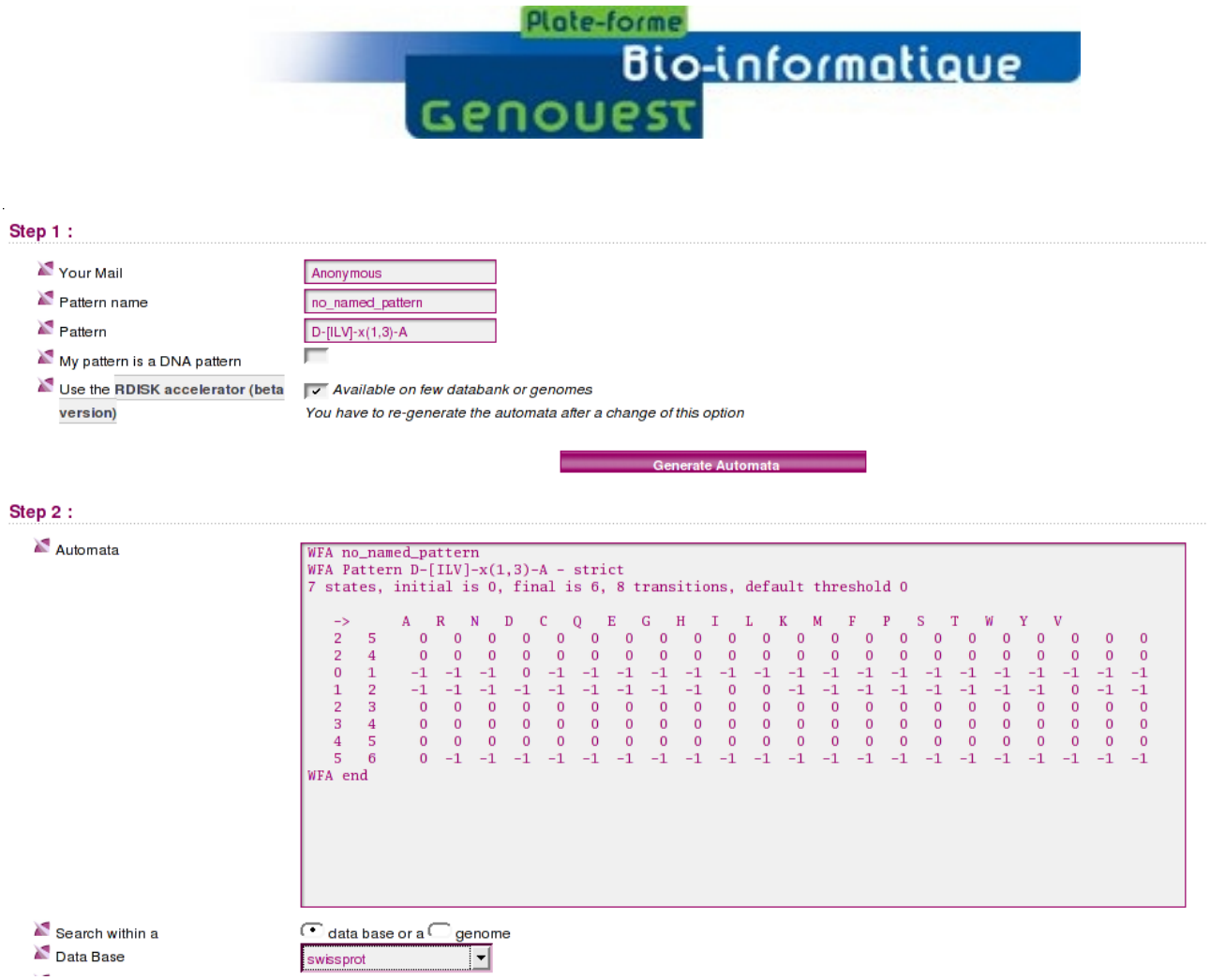


Plate-forme
Bio-informatique
Genouest
Step 1 :
Your Mail
Anonymous
Pattern name
no_named_pattern
Pattern
D-[ILV]-x(1,3)-A
My pattern is a DNA pattern
Use the RDISK accelerator (beta version)
Available on few databank or genomes
You have to re-generate the automata after a change of this option
Generate Automata
Step 2 :
Automata
WFA no_named_pattern
WFA Pattern D-[ILV]-x(1,3)-A - strict
7 states, initial is 0, final is 6, 8 transitions, default threshold 0
-> A R N D C Q E G H I L K M F P S T W Y V
2 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 1 -1 -1 -1 0 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
1 2 -1 -1 -1 -1 -1 -1 -1 -1 -1 0 0 -1 -1 -1 -1 -1 -1 -1 -1 0 -1 -1
2 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
3 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5 6 0 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
WFA end
Search within a
Data Base
data base or a genome
swissprot

Illustration 1: example of input data filled in the Web interface

It's possible for the user to modify the automata (chapter 4).

We could choose de research the pattern in a data bank -illustration 2).

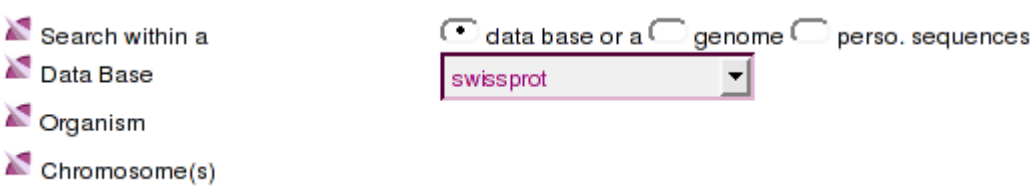


Illustration 2: an other example of input data filled in the Web interface : search within a data base.

We could choose de research the pattern in a genome (so in that case you must to select an organism and one or several chromosoms – to select several chromosoms with the touch «uppercase»- and verify that the option « each sequence matched » is selected.) or a personnal data bank (illustration 3 & 4).

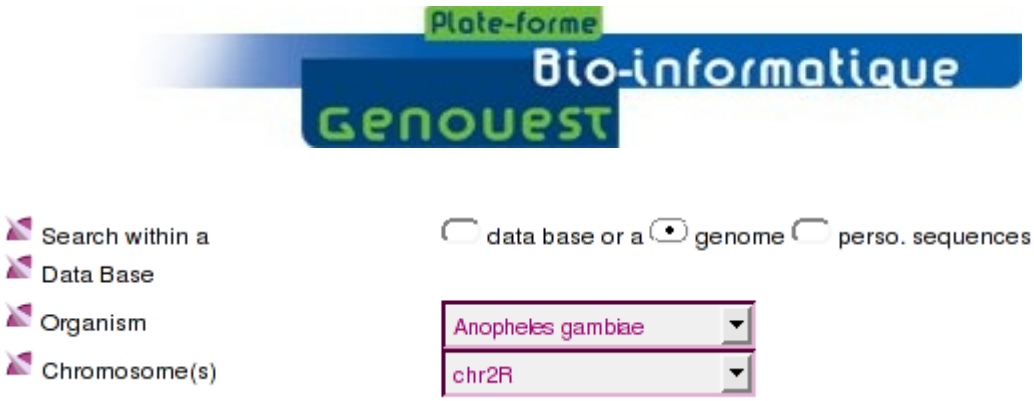


Illustration 3: an other example of input data filled in the Web interface : search within a genome.

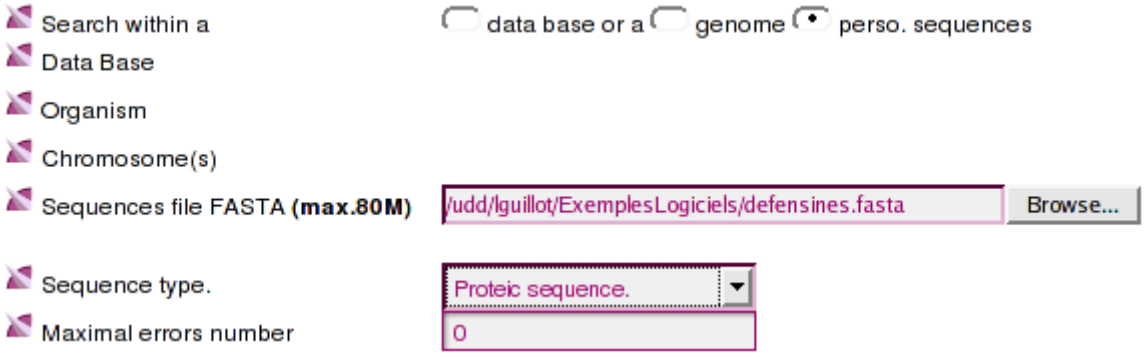
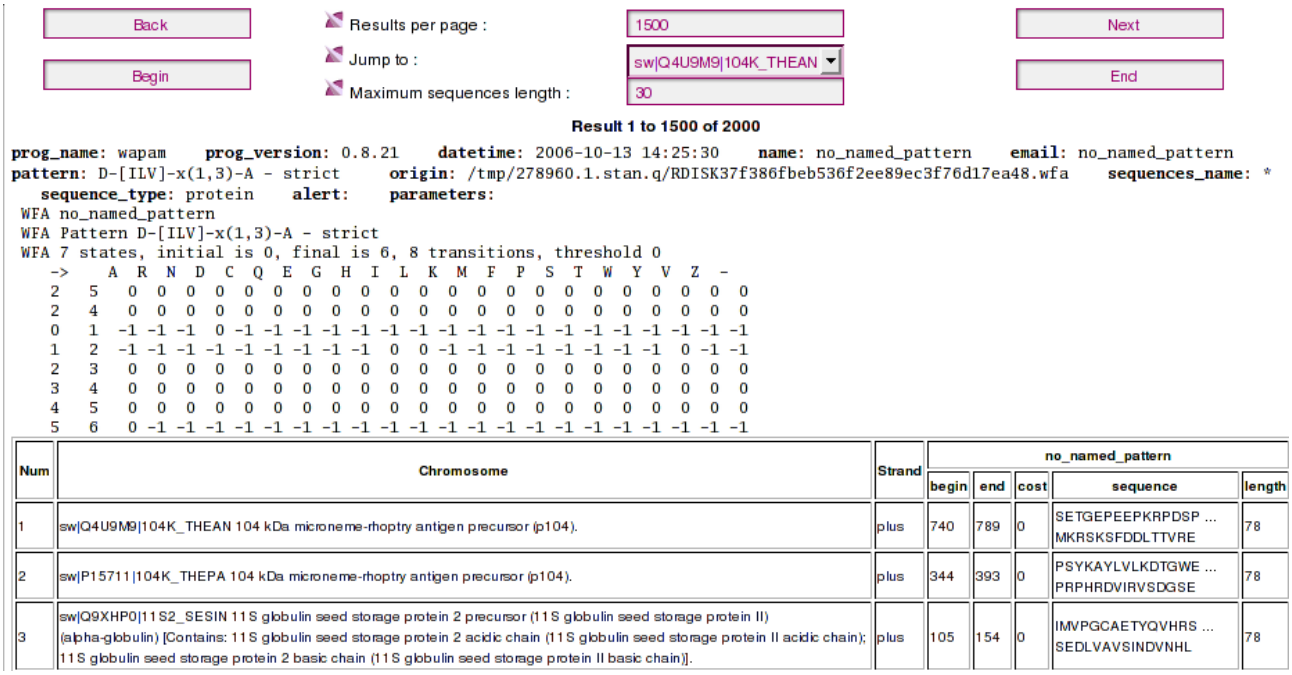


Illustration 4: an other example of input data filled in the Web interface : search within a personnal data bank.

A page of setting on standby post a progression indicator (illustration 5).

Illustration 5: progression of the compilation of the Rdisk (FPGA) processors before the sequences filtration.

The results are those printed illustration 6.



Back | Begin | Results per page : 1500 | Jump to : sw|Q4U9M9|104K_THEAN | Maximum sequences length : 30 | Next | End

Result 1 to 1500 of 2000

prog_name: wapam **prog_version:** 0.8.21 **datetime:** 2006-10-13 14:25:30 **name:** no_named_pattern **email:** no_named_pattern
pattern: D-[ILV]-x(1,3)-A - strict **origin:** /tmp/278960.1.stan.q/RDISK37f386fbeb536f2ee89ec3f76d17ea48.wfa **sequences_name:** *
sequence_type: protein **alert:** **parameters:**

```
WFA no_named_pattern
WFA Pattern D-[ILV]-x(1,3)-A - strict
WFA 7 states, initial is 0, final is 6, 8 transitions, threshold 0
   ->    A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  Z  -
   2   5   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
   2   4   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
   0   1  -1 -1 -1  0 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
   1   2  -1 -1 -1 -1 -1 -1 -1 -1 -1  0  0 -1 -1 -1 -1 -1 -1 -1 -1  0 -1 -1
   2   3   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
   3   4   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
   4   5   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
   5   6   0 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
```

Num	Chromosome	Strand	no_named_pattern				
			begin	end	cost	sequence	length
1	sw|Q4U9M9|104K_THEAN 104 kDa microneme-rhoptry antigen precursor (p104).	plus	740	789	0	SETGEPEEPKRPDSP ... MKRSKSFDDLTTVRE	78
2	sw|P15711|104K_THEPA 104 kDa microneme-rhoptry antigen precursor (p104).	plus	344	393	0	PSYKAYLVLKDTGWE ... PRPHRDVIRVSDGSE	78
3	sw|Q9XHP0|11S2_SESIN 11S globulin seed storage protein 2 precursor (11S globulin seed storage protein II) (alpha-globulin) [Contains: 11S globulin seed storage protein 2 acidic chain (11S globulin seed storage protein II acidic chain); 11S globulin seed storage protein 2 basic chain (11S globulin seed storage protein II basic chain)].	plus	105	154	0	IMVPGCAETYQVHRS ... SEDLVAVSINDVNHL	78

Illustration 6 : the results for the HTML example

Without Rdisk

The data seizures are the same as those in illustration 1,

Just don't select the checkbox Rdisk. The Illustration 7, shows the jobs number on standby on genocluster. The request is placed in the waiting jobs before its execution on one of the cluster nodes.

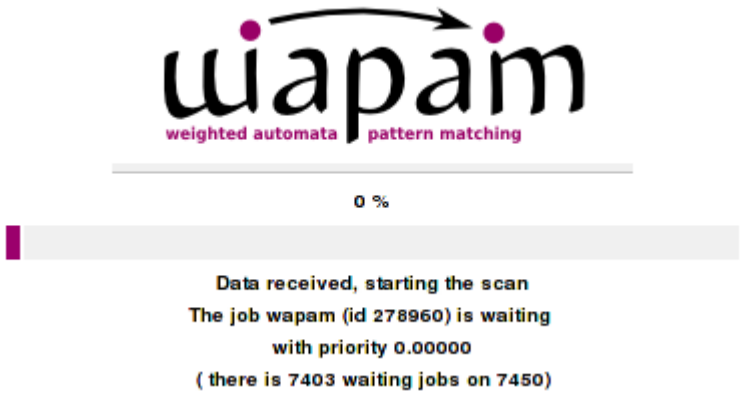

wapam
weighted automata pattern matching
0 %
Data received, starting the scan
The job wapam (id 278960) is waiting
with priority 0.00000
(there is 7403 waiting jobs on 7450)

Illustration 7: positioning of the job launched in the genocluster tasks queue.

The resultats are the same as those shown in the Illustration 6.

3 Input and output Data

Input data

The parameters to fill the form are as follows:

- Give your **email** est optional but advised. Certain research can be rather long, you are thus likely to close your navigator and to lose the url on the result page. In any case your result file is saved 5 days on our server.
- The **pattern name** is optionnal too. It allows you to distinct yours requests when you execute successive requests.
- If your pattern is a **nucleic** pattern you must to specify it.
- Choose to use **Rdisk** or not. The specialized machine Rdisk allows you to accelerate your research (see
- Choisir d'utiliser **Rdisk** ou non. La machine spécialisée Rdisk permet d'accélérer les calculs (see below). It 's a prototype of research that could be often out of order.
- Define the target sequences. The platform La plate-forme places approximately 200 genomes and 20 data banks at user's disposal. Genomes and data banks could be added, upon request (webmaster@genouest.org). If you use Rdisk the choice is limited but it could be supplemented too upon request. You can search too within your personnal sequences data bank.
- Choose the result type : each pattern that match in the sequences (« **each match** ») or each sequence that match with the pattern **« each sequence matched »)**. Usually, you choose « eatch match » (particulary if the research is made in a genome).

Form « Multiple Lauch »

Accessible by a bond that is in top on the left form. It allows user to repeatedly launch Wapam on a number of patterns (several patterns in a textfile but not use the Word format). The other input parameters are identical. In this case of use:

- The automata are not modifiable manually.
- The results are sonly send by email. Upon request, one mail by result or a single mail . In that last case, the results are written in a single file.

To have more informations on the launching of a whole of patterns or to set up a treatment with

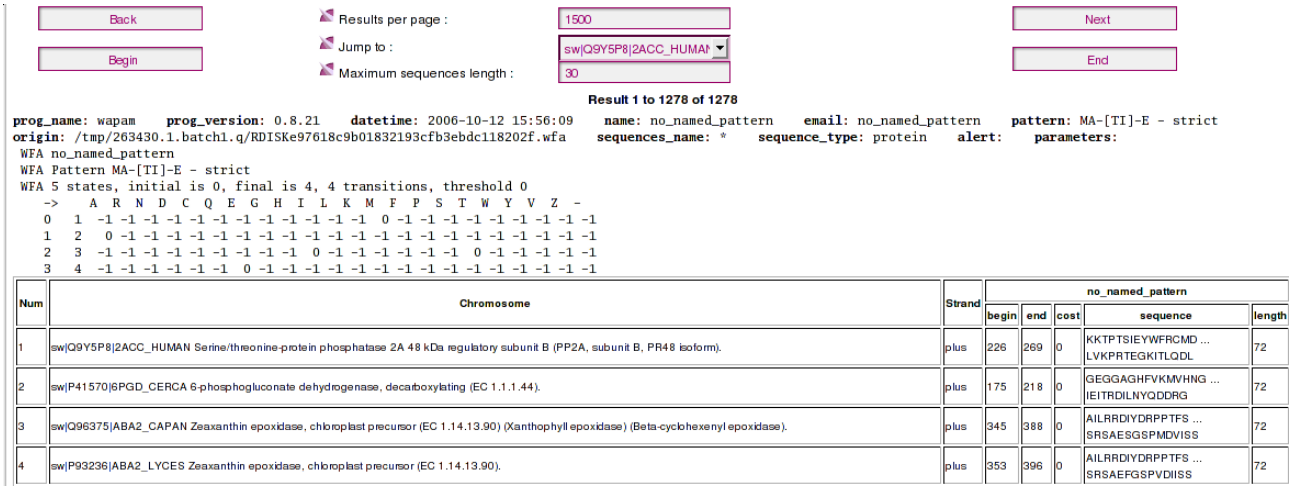
many patterns, contact webmaster@genouest.org.

Ouput data

The 3 formats of results description contain the same data exactly but they are presented differently.

Web Format (HTML)

The HTML format allows user to visualize the data in a table in your Internet explorer (illustration 8).



Back | Begin | Results per page : 1500 | Jump to : sw|Q9Y5P8|2ACC_HUMAN | Maximum sequences length : 30 | Next | End

Result 1 to 1278 of 1278

prog_name: wapam **prog_version:** 0.8.21 **datetime:** 2006-10-12 15:56:09 **name:** no_named_pattern **email:** no_named_pattern **pattern:** MA-[TI]-E - strict
origin: /tmp/263430.1.batch1.q/RDISKe97618c9b01832193cfb3ebdc118202f.wfa **sequences_name:** * **sequence_type:** protein **alert:** **parameters:**

```
 WFA no_named_pattern
 WFA Pattern MA-[TI]-E - strict
 WFA 5 states, initial is 0, final is 4, 4 transitions, threshold 0
    ->    A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  Z  -
    0   1  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1  0 -1 -1 -1 -1 -1 -1 -1 -1 -1
    1   2   0 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
    2   3  -1 -1 -1 -1 -1 -1 -1 -1 -1  0 -1 -1 -1 -1 -1 -1  0 -1 -1 -1 -1
    3   4  -1 -1 -1 -1 -1 -1 -1  0 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
```

Num	Chromosome	Strand	no_named_pattern				
			begin	end	cost	sequence	length
1	sw|Q9Y5P8|2ACC_HUMAN Serine/threonine-protein phosphatase 2A 48 kDa regulatory subunit B (PP2A, subunit B, PR48 isoform).	plus	226	269	0	KKTPTSIEYWFRCMD ... LVKPRTEGKITLQDL	72
2	sw|P41570|6PGD_CERCA 6-phosphogluconate dehydrogenase, decarboxylating (EC 1.1.1.44).	plus	175	218	0	GEGGAGHFVKMVHNG ... IEITRDILNYQDDRG	72
3	sw|Q96375|ABA2_CAPAN Zeaxanthin epoxidase, chloroplast precursor (EC 1.14.13.90) (Xanthophyll epoxidase) (Beta-cyclohexenyl epoxidase).	plus	345	388	0	AILRRDIYDRPPTFS ... SRSAESGSPMDVISS	72
4	sw|P93236|ABA2_LYCES Zeaxanthin epoxidase, chloroplast precursor (EC 1.14.13.90).	plus	353	396	0	AILRRDIYDRPPTFS ... SRSAEFGSPVDIISS	72

Illustration 8: example of the output data in HTML format with the option « each sequence matched»

The number of results printed on a page can be given by filling the field text “Result per pages” in top of the page (by defect 1500). The recovered data (illustration 6) are:

- The **chromosome or the sequence name**. You can go to the chromosome or the sequence which interests you by clicking at once “jump to” in top of the page.
- The **strand** (for the moment research is done only on the strand plus)
- The **position of beginning and the position of end** of the sequence printed in the results (and not those of the pattern).
- The **cost** or error number

- The **sequence** which you can select the length of posting in maximum field text the "sequences length" in top of the page (by defect 30).
- The **real length** of the part of the printed sequence.

1996	sw|P13671|CO6_HUMAN Complement component C6 precursor.	plus	441	481	0	ISLIRGGRSEYGAAL ... RNIPCAVTKRNNLRK	69
1997	sw|P13671|CO6_HUMAN Complement component C6 precursor.	plus	553	593	0	EKQSPDYKSNAVDGQ ... QEEDCTFSIMENNGQ	69

Illustration 9: example of results with the option "each match". In this sequence, the pattern appears twice at the positions 481 and 593.

XML Format

XML Format is a standard format (illustration 10), allowing to save data so that it can be read again easily by an human or a program. You could use it if you wish automatically to treat the data by a script that you would write yourself. Actually, the Web format is produced starting from XML format.

```
<?xml version="1.0" encoding="UTF-8" ?>
<result  prog_name="wapam"  prog_version="0.8.21"  datetime="2006-10-12 15:56:09"  name="no_named_pattern"  email="no_named_pattern"  pattern="MA-[TI]-E - stri
  <occurrence sequence="sw|Q9Y5P8|2ACC_HUMAN Serine/threonine-protein phosphatase 2A 48 kDa regulatory subunit B (PP2A, subunit B, PR48 isoform)." complement="
    <pattern name="no_named_pattern" begin="226" end="269" cost="0"  >
      <sequence type="protein"  >
        KKTPTSIEYWFRCMDLDGDGALSMFELEYFYEEQCRRLDSMAIEALPFQDCLCQMLDLVKPRTEGKITLQDL
      </sequence>
    </pattern>
  </occurrence>
  <occurrence sequence="sw|P41570|6PGD_CERCA 6-phosphogluconate dehydrogenase, decarboxylating (EC 1.1.1.44)." complement="plus"  >
    <pattern name="no_named_pattern" begin="175" end="218" cost="0"  >
      <sequence type="protein"  >
```

Illustration 10: example of the output data in XML format

CSV Format

Format CVS (illustration 11) allows the user to import the data in any software like Excel or Open Office.Calc. It is also translated starting from XML format. CSV Format used by WAPAM is as follows:

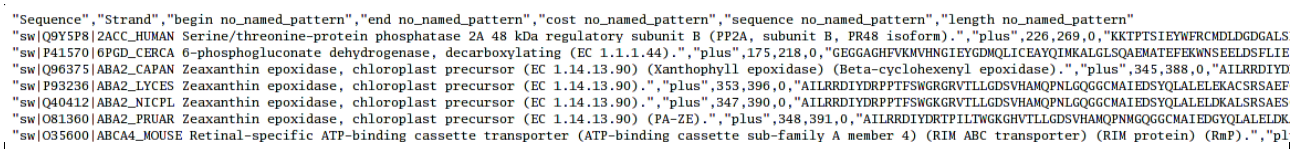
- the separator of fields is the comma,
- the separator of text is the quotation mark.

To import a CVS document in Excel :

1. On the WAPAM Web interface, click on the right button of the mouse on the bond 'Description of the results to format CSV ", finally click on 'Recording the target of the bond

under

2. In Excel : File/Open
3. Select "all" in 'file type'
4. Select the file type CVS adn submit
5. Select all the column A
6. In the menu "Data" select "Convert"
7. Choose option "délimited" and press on "following"
8. Indicate the separator : comma and the indicator of text : the quotation mark
9. Click on finish
10. You just have to format your table as good seems to you.



```
"Sequence","Strand","begin no_named_pattern","end no_named_pattern","cost no_named_pattern","sequence no_named_pattern","length no_named_pattern"
"sw|Q9Y5P8|2ACC_HUMAN Serine/threonine-protein phosphatase 2A 48 kDa regulatory subunit B (PP2A, subunit B, PR48 isoform).","plus",226,269,0,"KKTPTSIEYWFRCMDLDGDGALS
"sw|P41570|6PGD_CERCA 6-phosphogluconate dehydrogenase, decarboxylating (EC 1.1.1.44).","plus",175,218,0,"GEGGAGHFVKMVHNGIEYGDMQLICEAYQIMKALGLSQAEMATEFEKWNSEELDSFLIE
"sw|Q96375|ABA2_CAPAN Zeaxanthin epoxidase, chloroplast precursor (EC 1.14.13.90) (Xanthophyll epoxidase) (Beta-cyclohexenyl epoxidase).","plus",345,388,0,"AILRRDIYD
"sw|P93236|ABA2_LYCES Zeaxanthin epoxidase, chloroplast precursor (EC 1.14.13.90).","plus",353,396,0,"AILRRDIYDRPPTFSWGRGRVTLLGDSVHAMQPNLGQGGCMAIEDSYQLALELEKACSRSAEF
"sw|Q40412|ABA2_NICPL Zeaxanthin epoxidase, chloroplast precursor (EC 1.14.13.90).","plus",347,390,0,"AILRRDIYDRPPTFSWGKGRVTLLGDSVHAMQPNLGQGGCMAIEDSYQLALELDKALSRSAES
"sw|O81360|ABA2_PRUAR Zeaxanthin epoxidase, chloroplast precursor (EC 1.14.13.90) (PA-ZE).","plus",348,391,0,"AILRRDIYDRTPILTWGKGHVTLLGDSVHAMQPNMGQGGCMAIEDGYQLALELDK
"sw|O35600|ABCA4_MOUSE Retinal-specific ATP-binding cassette transporter (ATP-binding cassette sub-family A member 4) (RIM ABC transporter) (RIM protein) (RmP).","pl
```

Illustration 11 : example of the output data in CVS format

Notice about the number of results

We limited the Wapam results number (by genocluster: 2000/by Rdisk: 500). Indeed, a request with a too great number of results appears not easily interpretable: it is then preferable that the user biologist refine his research. It is however possible to increase these thresholds by contacting webmaster@genouest.org.

4 Some details about Wapam functionalities

Weighted automata (WFA)

An automata characterizing a pattern will be represented by the whole of the positions of the pattern, connected between them by transitions (illustration 12). The automata is weighted, i.e. that each transition is labelled by a letter which can be read according to the alphabet of the sequence (nucleic or proteic bases) and by a weight.

The sequence gradually “goes along” the automata, and, with each position, the weight of its transition is added to the score. This weight reflects the adequacy of the target sequence part (banks or genome) with the letter read with this position in the pattern. By defect this weight is equal to -1 if the letter is not the same one (substitution) and to 0 if it is the same one.

The reason is recognized when the final state is active with a score higher or equal to the score or fixed threshold of error. For example if an error is tolerated the threshold will be equal to -1.

On l'illustration 12 presenting an example of weighted automata, each round is a state, each arrow is a transition.

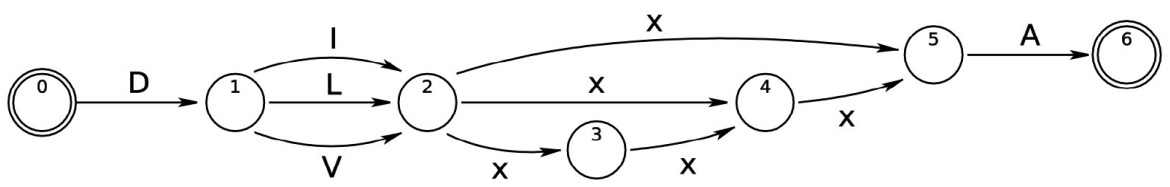


Illustration 12 : *a wheighted autamata of D-[ILV]-x(1,3)-A.*

The automatas used by Wapam are in the following form (illustration 13). For example, if the portion of sequence which goes along the automatq go from state 0 to 1 by reading D the cost is 0 if not the cost is -1.

```
7 states, initial is 0, final is 6, 8 transitions, default threshold 0

  ->      A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V
  2   5   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  2   4   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  0   1  -1  -1  -1   0  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1
  1   2  -1  -1  -1  -1  -1  -1  -1  -1  -1   0   0  -1  -1  -1  -1  -1  -1  -1   0  -1  -1
  2   3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  3   4   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  4   5   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  5   6   0  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1
```

Illustration 13 : example of automata representing a Prosite pattern D-[ILV]-x(1,3)-A.

The weights can be more general than the simple calculation “0/-1”; it is possible to modify the automata manually. For example the substitution of D by NR, R or A in first position can cost -3 instead of -1 (Illustration 14).

->		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V		
2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	-3	-3	-3	2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	2	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1
2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	6	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

Illustration 14 : e*xample of automata modified* .

The platform has other tools to generate waighted automatas (generation of weights “to the BLOSSUM”, use of matrices weights/position PWM…) Contact webmaster@genouest.org for questions on this subject.

Wapam and Wapam/Rdisk

Wapam can be used in two ways (Illustration 15) either it is launched on genocluster (like all the other softwares of the platform) and research is done on a node of the cluster, or it is coupled with Rdisk which parallels research on a whole of charts.

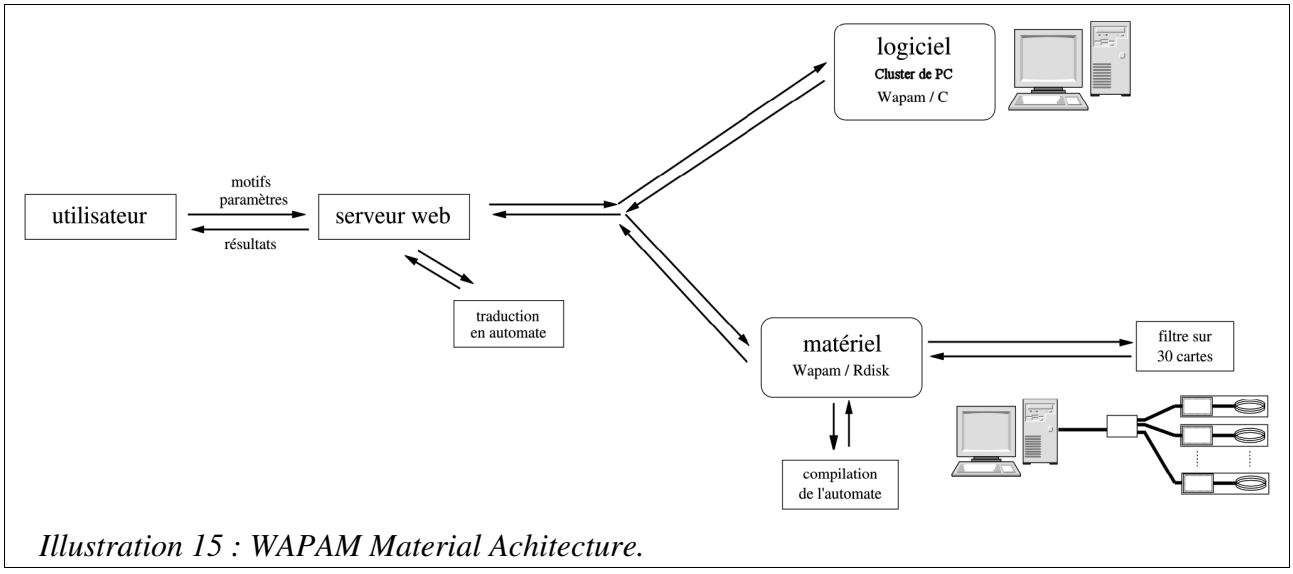


Illustration 15 : WAPAM Material Achitecture.

Rdisk is a specialized architecture made up of several tens of charts (currently 31). Each chart contains a reconfigurable processor (FPGA) coupled to a hard disk. The weighted automatas are directly wire up the FPGA, which allows a simultaneous evaluation of the states.

This wiring uses as many material elements as states transitions in the automata. The processors

have a surface being able to wire up automatas having until a hundred transitions. The 31 charts share the courses of the bank or the genome (1/31th by chart). The whole of the Rdisk prototype was conceived to filter the data bases quickly, the hard disks are directly connected to processors FPGA.

Rdisk is a prototype of research, so it is not always in operating state. If you need intensive calculations in search for patterns, contact the platform (webmaster@genouest.org) so that we set up an adapted treatment of your data or your patterns.

Performances

The illustration16 presents a comparison of the pattern search times between the Wapam software implementation and the Wapam/Rdisk material acceleration (average on 50 patterns taken by chance among a whole of 3331 patterns). Not to overload the server, the research can be stopped as soon as there is more than one certain number of results (hitch-hiking).

In all the cases, a research with Wapam with or without errors takes the same execution time. On the software version, the execution time is linear compared to the size of the automata (and thus of the pattern). With Wapam/Rdisk, all the patterns are treated in same time (as long as they are accepted by Rdisk, i.e. as long as there is not more than one hundred of transitions).

	Wapam software	***Wapam + autostop 2000***	***Wapam/Rdisk***	***Wapam/Rdisk + precompilation***
1 pattern	2605 s	2003 s	72 s	23 s
3331 patterns	100 days*	77 days*	< 3days	< 1day

Illustration 16 : *comparison between the search times of pattern (*: estimates)*

The acceleration brought by Rdisk is even more important starting from the second launching, when the patterns were already compiled, because Wapam/Rdisk remembers the weighted automatas compiled previously. The modification of the error threshold doen't require a new compilation.

Specific needs

We are at your disposal (webmaster@genouest.org) to collaborate on particular tasks, such as for example:

- to add other data banks,
- to produce weighted automatas meeting particular aims,

- to set up on the cluster or Rdisk intensive calculations (great number of sequences, of patterns/ automatas, reiterated launchings, analyzes results…); we can finely parameterize Wapam to obtain the best computing times on your application,
- To allow you to access to Wapam by commands lines on genocluster

References

Thank you to quote the following reference in your work using Wapam.

Stéphane Guyetant, Mathieu Giraud, Ludovic L'Hours, Steven Derrien, Stéphane Rubini, Dominique Lavenier, and Frédéric Raimbault. Cluster of re-configurable nodes for scanning large genomic banks. Parallel Computing, 31(1):73-96, 2005.